DOCUMENT RESUME

ED 409 331 TM 026 773

AUTHOR Yang, Wen-Ling
TITLE Validity Issues in Cross-national Relational Analyses: A Meta-Analytic Approach to Perceived Gender Differences on Mathematics Learning.
PUB DATE Mar 97
NOTE 55p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cross Cultural Studies; Effect Size; Foreign Countries; Grade 7; Grade 8; International Education; *International Studies; Junior High Schools; *Mathematics; *Meta Analysis; Regression (Statistics); *Sex Differences; Student Attitudes; *Validity
IDENTIFIERS *Third International Mathematics and Science Study

ABSTRACT

International comparisons in educational research can be difficult to accomplish because the findings of individual countries may not be comparable due to study design or inherent country features that cannot be manipulated. Quantitative meta-analysis techniques have great potential in improving international comparisons. In this paper, participant countries/regions in an international study were treated as study populations, and meta-analytic techniques were used to synthesize study outcomes across countries. Homogeneity tests were conducted to determine whether there was a common population parameter across countries, outliers were identified empirically, moderator effects die to country characteristics were studied, and homogeneous country outcomes were combined by a variance-weighting method to yield an optimal parameter estimate. The study of interest in this paper was gender differences in students' perceptions about whether girls or boys will do better in mathematics. Data are a subset of data from the Third International Mathematics and Science Study (TIMSS), in which seventh- and eighth-graders from 25 countries participated. Multiple regression models were used to country-level analyses, and effect sizes were computed. Overall, the meta-analytic techniques were satisfactory in analyzing the TIMSS data, illustrating the potential of meta analysis in improving the validity of international comparison studies. Meta analysis appears to be capable of detecting substantial differences in country outcomes and effective in offering strategies to deal with the situation. Appendixes present tables of coding schemes and outcomes of moderators, independent variables and questionnaire items, and a summary of meta analytic results. (Contains 7 tables, 8 figures, and 20 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 409 331

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Wen-Ling Yang
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Validity Issues in Cross-national Relational Analyses: A Meta-Analytic Approach to Perceived Gender Differences on Mathematics Learning

Wen-Ling Yang

Michigan State University

Paper presented at the 1997 AERA Annual Meeting in Chicago

I gratefully acknowledge the assistance of Dr. Betsy J. Becker for her comments and editing on a draft of this article, and the TIMSS US National Research Center for the permission of using its data.

Inquiries concerning this paper should be sent to Wen-Ling Yang, who is now at the Department of Counseling, Educational Psychology, and Special Education, Michigan State University, East Lansing, MI 48824.

TM026773

Introduction

One important purpose of international comparison is to compare educational phenomena of interest across countries to determine the degree of similarity or disparity. Ideally, a more representative global picture will emerge after combining the results from homogeneous countries and contrasting the discrepancies among heterogeneous countries. However, these tasks are often difficult to accomplish, for sometimes the findings of individual countries are not compatible due to different study designs or inherent country features that cannot be manipulated. Moreover, it may be desirable not to incorporate artificial manipulations so that genuine variations across countries can by studied under natural contexts.

Inconsistent findings in comparative studies, therefore, either reflect true unique country characteristics or can be explained by the variations in study designs. The trade-off between the desire for compatibility of data and the need to study naturally occurring phenomena has complicated interpretations of the comparative study results. Conventional qualitative review methods are judgment-based and usually fail to provide statistically justifiable explanations for the similarities or differences among countries, and nor do these methods offer sensible strategies for summarizing incompatible country outcomes due to different study or inherent country features.

Quantitative meta-analysis methods have great potentials in improving international comparisons. In this paper, participant countries/regions in an international study were treated as study populations and meta-analytic techniques were applied to synthesize study outcomes across countries. Homogeneity tests were conducted to determine whether there was a common population parameter acrqss countries, outliers were empirically identified, moderator effects due to important country characteristics were studied to account for between-countries differences, and homogeneous country outcomes were combined by variance-weighting method to yield an optimal parameter estimate.

The cross-national study outcome of interest in this paper is gender difference in students' perceptions about whether girls or boys will do better on math. For the country-level studies, multiple regression models were developed and the overall model efficacy and the unique effect due to gender were studied for all the countries/regions. Several related summary statistics were also obtained and meta-analyzed. By exploring the effectiveness of meta-analytic approaches in analyzing complex cross-national data, hopefully, the quantitative synthesis of international study results will be improved. Incorporated with careful qualitative considerations of individual country characteristics, a more realistic picture accounting for between-country disparities will emerge from meta-analysis outcomes.

The design, method, and results of primary studies are summarized in the next section, followed by a brief description of meta-analytic techniques used. Then, meta-analysis results are presented and discussed.

Primary Studies

Gender differences in mathematical outcomes are fairly well established (Maccoby & Jacklin, 1974). It is also found that gender plays an important role in students' perceptions and beliefs in learning mathematics (Fennema, 1974; Fennema & Sherman, 1977), which is possibly due to differential socialization (Maccoby & Jacklin, 1974; Mayer, 1987). In studies of motivational factors, such as self-efficacy and aspiration level, it is found that student's self-perception and belief in learning correlate with student's learning achievement (Bandura, 1982; Bandura & Schunk, 1981; Hermans, 1970; Norwich, 1986; Schunk, 1981; Schunk, 1988). One plausible contribution to gender differences in students' mathematics achievement, therefore, is students' beliefs in whether girls or boys will do better on math. It is important to study the possible causes of the gender difference in students' beliefs about math learning.

The goal at the primary-study level of this research was to explore, under the context of individual countries/regions, whether and how student gender associated with student belief in which gender group would do better on

mathematics. Several other independent variables were also included in multiple regression analyses to yield useful predictive models for students' beliefs.

The data analyzed in this study are a subset of data from the field-trial version of the student questionnaires of the Third International Mathematics and Science Study (TIMSS). General features of the data and operational definitions of the variables are outlined below, followed by brief descriptions of the design, method, and results of the primary studies.

Description of Primary Data

Information about high-school students' perceived gender differences on mathematics learning and potential influential factors such as gender, self-efficacy, and parents' expectations were collected by the national research centers of the countries/regions participating in the TIMSS field-trial student survey.

Twenty-five individual countries/regions (or studies) were included at the primary level. In most cases, data were collected at the country level. However, separate data were collected for Flemish Belgium and French Belgium. A closer examination revealed that the two regions differed to a significant degree in terms of the nature of the student samples, the means and standard deviations of the dependent variables, as well as most of the analysis results. Therefore, the two regional datasets were included in this study, instead of the combined national dataset of Belgium. The dataset of the Canadian province of Ontario was excluded because it overlapped with the overall Canada dataset.

According to the design of the trial, national representative samples were drawn from the populations of students at the 7th and the 8th grades from all the countries/regions. The ages of the students, according to the TIMSS definition, should have been between 13 and 15 when the data were collected. Nevertheless, a frequency analysis on student age showed a considerably wider range. The bewildering range, showing students from the ages of 10 to 18, might reflect differences in the educational systems of various countries. Fortunately, the overall variance in age was not too big. More than 95% of the

students were clustered at the categories of age 13 to 15. Therefore, the small proportions of students at the two ends were neglectable and the populations were treated as homogeneous in terms of both grade and age.

Because the data was collected for use in the final revision of the TIMSS questionnaires, it was not guaranteed that the samples drawn on this basis were truly nationally representative. Since the sample sizes were all relatively small, these samples were likely to be convenient and hence potentially biased. Though the research centers were instructed to draw randomly representative samples, intact classes might have been used. Further, the classes were more likely to be drawn from less distant urban areas, where the students were probably more serious and competitive about their learning and the general residents were more enthusiastic about education.

The potential sample bias discussed above not only posed an immediate threat to the interpretation and generalization of the study results, it also raised questions on the validity of using commonly known country characteristics as moderator variables in meta-analysis in this study. To cope with such threat, instead of using national statistics based on representative samples such as economical developmental level or index of modernity, this study coded relevant information from the same dataset to form country-characteristic (or study-characteristic) variables. The coding schemes are summarized in Appendix A.

Design of Primary Studies

Although the questionnaire items included in the analysis were originally written for the field trial, all of the items were later included in the final version of the TIMSS questionnaire. Therefore, they should all have reasonable validity. Operational definitions of the outcomes and the independent variables of the multiple regression models are presented below.

Outcome Variable

The outcome variable in the regression model for all countries/regions was a measure of the construct delineating student's perception about whether girls or boys would do better on mathematics. The measure was a composite

score on a set of eight items in the TIMSS student questionnaire. Students indicated their beliefs in mathematics learning on a 5-point Likert scale with the following categories: Boys, Boys more than girls, Boys and girls the same, Girls more than boys, and Girls. The questions were as follows:

Who do you think is more likely to:
- be better at mathematics?
- be interested in a career that uses mathematics?
- be more likely to solve a difficult mathematics problems?
- have a natural talent for mathematics?
- be comfortable asking questions in mathematics class?
- be encouraged by their mathematics teacher?
- be interested in mathematics?
- worry about how well they are doing in mathematics?

A composite score was formed empirically by summing up the eight items for each of the countries/regions, although this led to a trade-off in that not all the variance-covariance matrices across countries/regions would be maximized. Alternatively, the composite score could have been formed by using principle-component weights to maximize variance-covariance. Nevertheless, the sets of weights were not identical for all countries/regions, which would make cross-country comparisons impossible. Because the results of principle component analyses showed similar patterns of factor loadings for different countries/regions (the signs were roughly the same and the magnitudes were all close to one), items were weighted equally and summed up for each country/region to ensure cross-countries comparability.

Independent Variables

Several independent variables, other than gender, were included to explore the usefulness of the multiple regression models. An inherent limit on the inclusion of independent variables was that only those appearing in the TIMSS questionnaire were available. Given the limited availability of theoretical relevant independent variables in the TIMSS questionnaire, the selected independent variables were not expected to explain most (or even much) of the variation in the outcome variable.

The seven independent variables included in the primary study are (a) student gender, (b) general educational aspiration of student, (c) achievement

attribution of student, (d) student's general preference of mathematics, (e) parents' education, (f) perceived expectations of student on learning mathematics, and (g) student's self-efficacy. The measurement indices (component questionnaire items) of these variables were summarized in Appendix B. Generally, the correlations among the independent variables were low, with gender accounting for most of the variance in the perceived gender differences in mathematics learning.

Methods of Primary Studies

Analytic methods used for country-level analyses are summarized below.

Multiple Regression Models

For all of the countries/regions, multiple regression equations were formed to determine the usefulness of the seven independent variables and to test for the significance of gender alone. Incremental (partial) R^2 due to gender was computed and its significance was tested using an F-test (Kerlinger & Pedhazur, 1973). Usually, R^2 is adjusted for its degree of freedom (Shavelson, 1988). Nevertheless, to avoid getting negative values, the R^2s in this study were not statistically adjusted.

Computation of Effect Sizes

Effect sizes representing the gender effect without controlling for other important variables were also studied. The effect sizes, contrasting gender group differences, were computed and unbiased using the following formulas (Hedges & Olkin, 1985):

$$\text{Effect Size (d)} = (\text{Mean}_f - \text{Mean}_m) / S_{pooled}$$

$$\text{Unbiased Effect Size (t)} = J \times d;$$

where $J=1-(3/(4\times(n_f+n_m-2)-1))$, n_f and n_m are the sample sizes for the female and male groups respectively, Mean_f and Mean_m are the means for the female and male groups respectively, and S_{pooled} is the pooled standard deviation for the gender groups. The bias being corrected was due to small sample size.

Results of Primary Studies

Important findings and implications of the country-level analyses are summarized and discussed below.

Descriptive Statistics

Descriptive statistics needed for subsequent meta-analysis, including sample sizes and the means and standard deviations for gender groups, are summarized in Table 1. Roughly speaking, the sample sizes were balanced for the two gender groups across countries/regions, except for one region.

Interpretations of Effect Sizes

In this study, higher scores on perceived gender differences indicate a perception of female superiority in learning math and lower scores suggest a belief in male superiority. As a result, effect sizes in this study should be interpreted differently, depending on (a) their directions, and (b) the magnitudes of the group means. Table 2 presents a breakdown of effect sizes that helps reveal the nature of gender-group differences.

Major findings. From the cross-tabulation analysis, it was found:

- In nine of the 25 countries/regions, both male and female student thought that female students would do better in math, but female students perceived more superiority for female students.
- In only one country/region, both student groups thought that female students would do better in math, but male students perceived more superiority for female students.
- In five of the 25 countries/regions, both groups perceived that male students would do better, but female students tended to perceive less inferiority for female students.
- In two countries/regions only, both groups perceived that male students would do better, but female students tended to perceive more inferiority for female students.
- In eight of the 25 countries/regions, female students perceived female

Table 1

Descriptive Statistics of Primary Studies

Study ID (Country/Region)	n	n_f	n_m	sd_f	sd_m	m_f	m_m	Direction of group difference ($m_f - m_m$)	Classification* category
A	322	177	145	3.335 <	3.536	25.565	24.097	+	1
B	358	185	173	3.445 <	5.207	23.243	21.844	+	2
C	108	80	28	2.990 <	5.459	25.088	24.393	+	1
D	127	56	71	3.511	3.280	24.500	23.197	+	3
E	432	222	210	3.201 <	3.636	24.968	24.629	+	1
F	279	129	150	2.536 <	2.985	22.791	22.333	+	2
G	304	135	169	3.882 <	4.723	22.719	22.012	+	2
H	298	150	148	3.974 <	5.439	24.940	22.568	+	1
I	303	165	138	3.617 <	4.636	26.079	24.210	+	1
J	340	160	180	2.955 <	3.815	25.119	24.711	+	1
K	239	129	110	5.167	5.130	26.364	22.000	+	3
L	428	244	184	3.705 <	3.801	26.742	24.598	+	1
M	286	155	131	3.166 <	3.737	26.039	25.305	+	1
N	173	73	100	4.908 <	5.157	25.329	21.300	+	3
O	401	194	207	4.041 <	4.394	22.144	22.502	-	5
P	235	126	109	6.677 <	7.079	26.468	24.312	+	1
Q	213	105	108	3.082 <	4.576	24.552	21.778	+	3
R	283	123	160	3.042 <	3.888	23.163	23.663	-	5
S	60	36	24	3.840 <	4.498	23.667	21.667	+	2
T	310	175	135	4.754 <	5.938	25.303	22.911	+	3
U	301	154	147	5.759	5.709	26.020	21.374	+	3
V	236	116	120	4.262 <	4.759	24.198	22.258	+	3
W	396	193	203	4.219 <	5.564	23.005	21.744	+	2
X	211	104	107	3.266 <	4.485	24.394	24.542	-	4
Y	612	330	282	3.500 <	4.277	25.218	23.840	+	3

Note: 1. Range of score=(8,40) and middle score=24

2. *- To interpret the difference score, studies are classified by the following scheme:

If $m_f>24$, $m_m>24$, and $m_f-m_m>0$, then classification category=1

If $m_f<24$, $m_m<24$, and $m_f-m_m>0$, then classification category=2

If $m_f>24$, $m_m<24$, and $m_f-m_m>0$, then classification category=3

If $m_f>24$, $m_m>24$, and $m_f-m_m<0$, then classification category=4

If $m_f<24$, $m_m<24$, and $m_f-m_m<0$, then classification category=5

Table 2

Categorizing Effect Sizes

# of Effect Sizes / Group Perceptions / Direction of Effect Size	Both perceived female superity	Both perceived male superity	Opposite directions	
Positive (higher female scores)	9	5	8*	22
Negative (higher male scores)	1	2	0**	3
	10	7	8	25

Note: *- Female students thought female students would do better in math, whereas male student thought male students would do better.

**- Female students thought male students would do better in math, whereas male student thought female students would do better.

superiority on mathematics, but male students believed in male superiority.

It should be noted that a negative effect size may suggest that average female students perceive male superiority on math but male students believe in female superiority. However, such situation was not found in this study.

Implications for differential perceptions. Overall, females scored higher on perceived gender differences on math learning in 22 of the 25 countries/regions. All of the above findings seemed to suggest that female students generally believed that they could do better in math than male students (in 18 countries/regions), and even when they perceived a male superiority, female students tended to think themselves as less inferior than male students thought the female students would be (in 5 out of 7 countries/regions). But the findings also suggested that male students generally believed that they could do better in math than female students (in 15 countries/regions), and even when they perceived a female superiority, male students tended to think female students as less superior than male students themselves (9 out of the 10 countries/regions).

In addition, in ten of the 25 countries/regions, all students thought that female students would do better in mathematics; whereas in seven other countries/regions, all students thought that male students would do better in mathematics. The descriptive statistics in Table 1, however, showed that the perceptions of male students were generally more heterogeneous than the perceptions of the female students (22 out of 25 countries/regions).

The primitive review of the findings of country-level analyses led to somewhat disjointed conclusions and these conclusions were merely tentative. Beyond such review, meta-analysis should reach more plausible interpretations about differential perceptions of gender groups by taking into account the sample sizes and the group variances.

Overview of Various Summary Statistics

The summary statistics from the multiple regression analyses and incremental F-tests are presented in Table 3. They include the R^2 of the overall multiple regression model, the partial R^2 for gender, and the partial

Table 3

Summary Statistics of Primary Studies

Study ID (Country/ Region)	Model R^2 (7 predictors)	p value	Sig. (α=.05)	Model R^2 (6 predictors)	R^2_{change} (partial R^2_{gender})	F_{change}	p value of F_{Change}	Sig. (α=.05)	$\hat{\beta}_{gender}$	s.e. of $\hat{\beta}_{gender}$
A	0.060	0.007	*	0.025	0.035	11.605	0.001	*	-1.394	0.409
B	0.087	0.000	*	0.072	0.016	5.942	0.015	*	-1.129	0.463
C	0.030	0.878	n.s.	0.026	0.004	0.411	0.523	n.s.	-0.581	0.907
D	0.139	0.011	*	0.111	0.029	3.954	0.049	*	-1.225	0.616
E	0.011	0.679	n.s.	0.007	0.004	1.749	0.187	n.s.	-0.442	0.335
F	0.027	0.386	n.s.	0.024	0.003	0.763	0.383	n.s.	-0.309	0.354
G	0.026	0.339	n.s.	0.018	0.008	2.431	0.120	n.s.	-0.806	0.517
H	0.108	0.000	*	0.041	0.068	21.931	0.000	*	-2.589	0.553
I	0.075	0.002	*	0.027	0.048	15.214	0.000	*	-1.862	0.477
J	0.009	0.678	n.s.	0.006	0.004	1.277	0.259	n.s.	-0.425	0.376
K	0.161	0.000	*	0.012	0.150	41.181	0.000	*	-4.422	0.689
L	0.113	0.000	*	0.040	0.073	34.514	0.000	*	-2.128	0.362
M	0.040	0.123	n.s.	0.024	0.016	4.551	0.034	*	-0.886	0.415
N	0.145	0.000	*	0.025	0.120	23.197	0.000	*	-3.940	0.818
O	0.020	0.150	n.s.	0.018	0.002	0.906	0.342	n.s.	**0.405**	0.425
P	0.048	0.132	n.s.	0.025	0.023	5.411	0.021	*	-2.180	0.937
Q	0.139	0.000	*	0.024	0.116	27.551	0.000	*	-2.869	0.547
R	0.015	0.753	n.s.	0.007	0.008	2.221	0.137	n.s.	**0.668**	0.448
S	0.220	0.061	n.s.	0.201	0.019	1.286	0.262	n.s.	-1.279	1.128
T	0.075	0.001	*	0.024	0.051	16.714	0.000	*	-2.500	0.612
U	0.184	0.000	*	0.030	0.154	55.311	0.000	*	-4.890	0.658
V	0.049	0.115	n.s.	0.007	0.042	10.112	0.002	*	-1.918	0.603
W	0.046	0.011	*	0.032	0.014	5.562	0.019	*	-1.182	0.501
X	0.035	0.401	n.s.	0.035	0.000	0.019	0.891	n.s.	**0.076**	0.551
Y	0.044	0.000	*	0.017	0.028	17.435	0.000	*	-1.318	0.316

regression coefficient of the gender variable. The results of appropriate significance tests were also included. A significant test result for the multiple regression equations suggests that the predictor variables collectively accounted for a significant amount of variance in the outcome variable; a significant result for the partial R^2 indicates non-zero unique variance explained by gender.

d: unbiased effect size. The unbiased effect magnitudes, presented in Figure 1, showed some degree of variation in the effect sizes, though the values of R^2 were all relatively small. The unweighted mean of the effect sizes was about .34 and the standard deviation was .27. The minimum was -.14 and the maximum was .84.

R^2: proportion of variance explained. Figure 2 and Figure 3 display the R^2s from the 7-predictor regression model and the partial R^2s due to gender, respectively, sorted in descending order.

Generally, the R^2 values of the 7-predictor regression model were not big, ranging from .009 to .220. Most of the R^2s were smaller than .050 and the unweighted mean was .076. At the significance level of α=.05, 13 of the 25 R^2s of the regression models were significant.

As expected, the partial R^2s due to gender were also small, ranging from .0001 to .154. More than half of the partial R^2s were less than .020 and the unweighted mean was .041. At the significance level of α=.05, 16 of the 25 partial R^2s due to gender were significant. Thirteen of these 16 cases overlapped with the 13 cases found for the R^2 of the overall regression models.

$\hat{\beta}_{gender}$: partial regression coefficient. Partial regression coefficients of the gender variable and corresponding standard error estimates were obtained from the overall multiple regression equations. Overall, the values of the standard errors (ranging from .316 to 1.128, with a mean of 1.657) were not too high, compared to the magnitudes of the regression coefficients (ranging from .076 to 4.890, with a mean of .561).

Study ID (Country/Region)	Unbiased Effect Size $d=J^*[(m_f-m_m)/S_{pooled}]$
A	0.427
B	0.318
C	0.183
D	0.383
E	0.099
F	0.164
G	0.161
H	0.497
I	0.453
J	0.118
K	0.845
L	0.571
M	0.213
N	0.793
O	-0.085
P	0.313
Q	0.707
R	-0.141
S	0.480
T	0.450
U	0.808
V	0.428
W	0.254
X	-0.037
Y	0.355

(Unweighted) Mean= 0.350
(Unweighted) s.d.= 0.268

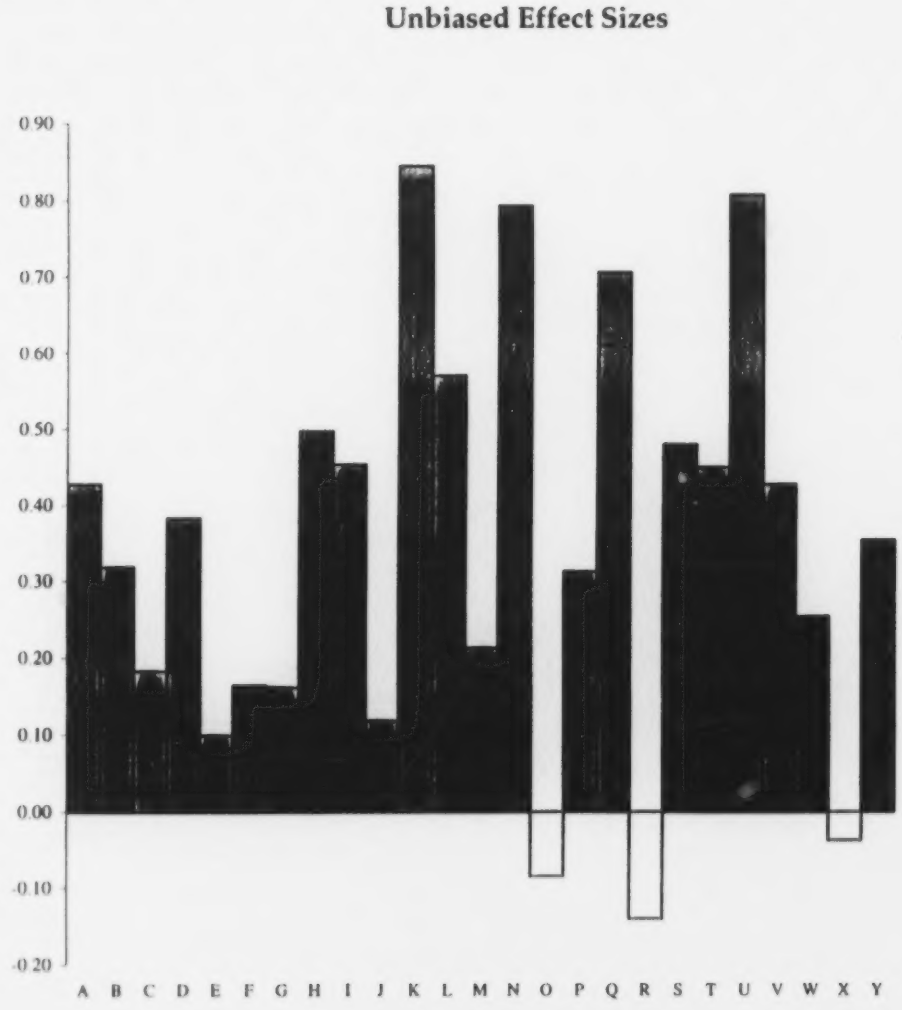

Figure 1. Unbiased effect sizes for perceived gender differences

Study ID (Country/ Region)	R^2
S	0.220
U	0.184
K	0.161
N	0.145
D	0.139
Q	0.139
L	0.113
H	0.108
B	0.087
I	0.075
T	0.075
A	0.060
V	0.049
P	0.048
W	0.046
Y	0.044
M	0.040
X	0.035
C	0.030
F	0.027
G	0.026
O	0.020
R	0.015
E	0.011
J	0.009

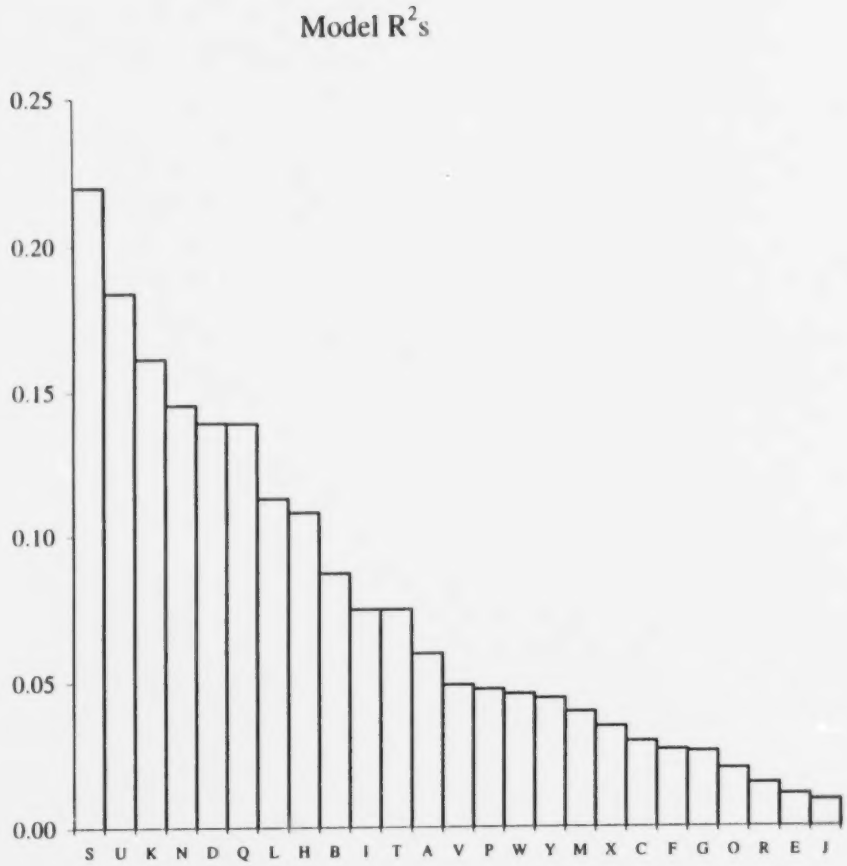

Figure 2. R^2 from multiple (7-predictor) regression model

Study ID (Country/Region)	Partial R^2_{gender}
U	0.154
K	0.150
N	0.120
Q	0.116
L	0.073
H	0.068
T	0.051
I	0.048
V	0.042
A	0.035
D	0.029
Y	0.028
P	0.023
S	0.019
M	0.016
B	0.016
W	0.014
G	0.008
R	0.008
E	0.004
C	0.004
J	0.004
F	0.003
O	0.002
X	0.000

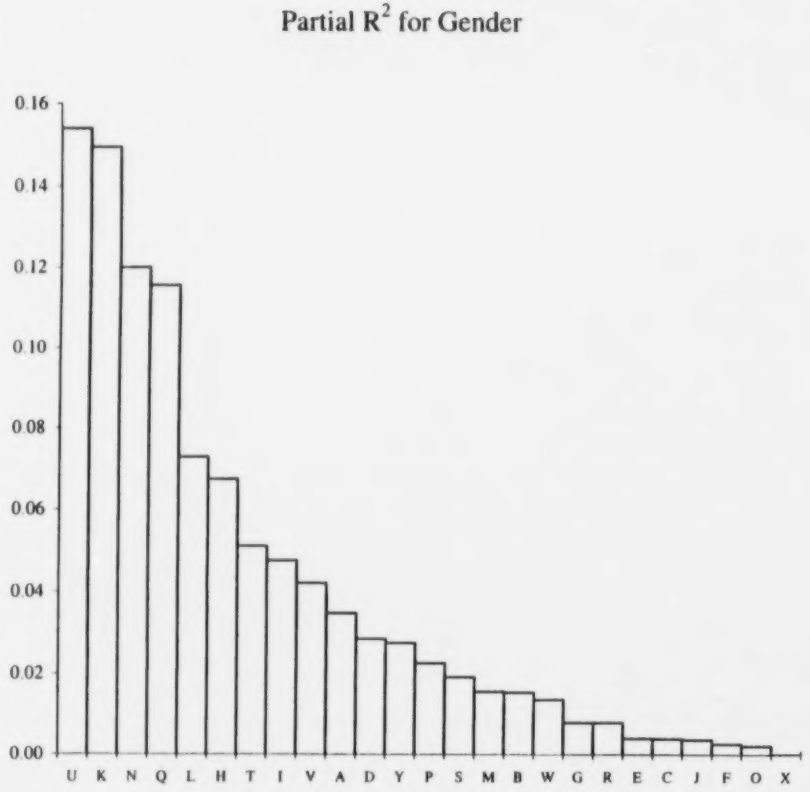

Figure 3. Partial R^2 for gender

Although the magnitude of a partial regression coefficient is influenced by the variance of the predictor variable (Shavelson, 1988), because the scale of the predictor variable was invariant across all the countries/regions, the regression coefficients were already comparable and there was no need to standardize the coefficients.

p value: observed exact probability. The p values for the tests of regression model fit, and the tests of the partial R^2, are jointly displayed in Figure 4. Generally, both sets of p values varied widely across countries/regions. However, most pairs of p values for individual countries were pretty similar. The exceptions were the first few countries shown in Figure 4, where the p values looked far apart.

Meta-Analysis

To study differential regression model fit across countries/ regions, and to compare the unique effects of gender on students' perceptions about whether girls or boys would do better on mathematics, various meta-analytical techniques were applied to the summary statistics from the 25 primary studies.

Methods

The following meta-analytical approaches were used for the synthesis of various study findings.

Homogeneity Test and Outlier Analysis

Homogeneity tests (Hedges & Olkin, 1985; Shadish & Haddock, 1994) were used to determine whether a common population parameter could represent the different countries/regions. If the countries/regions appeared heterogeneous, residual analyses (Hedges & Olkin, 1985) and analysis of moderators (Eagly & Wood, 1994) were conducted. Two moderators reflecting country biases were explored.

Explanatory Effect of Moderators

Two moderator variables were incorporated to account for variation among study outcomes of various countries/regions, and their significance were tested. The moderators were constructed to capture two relevant characteristics of countries/regions: (a) the general math achievement level of

Study ID (Country/Region)	p value for R^2_{total}	p value for partial R^2_{gender}
C	0.878	0.523
R	0.753	0.137
E	0.679	0.187
J	0.678	0.259
X	0.401	0.891
F	0.386	0.383
G	0.339	0.120
O	0.150	0.342
P	0.132	0.021
M	0.123	0.034
V	0.115	0.002
S	0.061	0.262
D	0.011	0.049
W	0.011	0.019
A	0.007	0.001
I	0.002	0.000
T	0.001	0.000
N	0.000	0.000
Y	0.000	0.000
B	0.000	0.015
H	0.000	0.000
K	0.000	0.000
L	0.000	0.000
U	0.000	0.000
Q	0.000	0.000

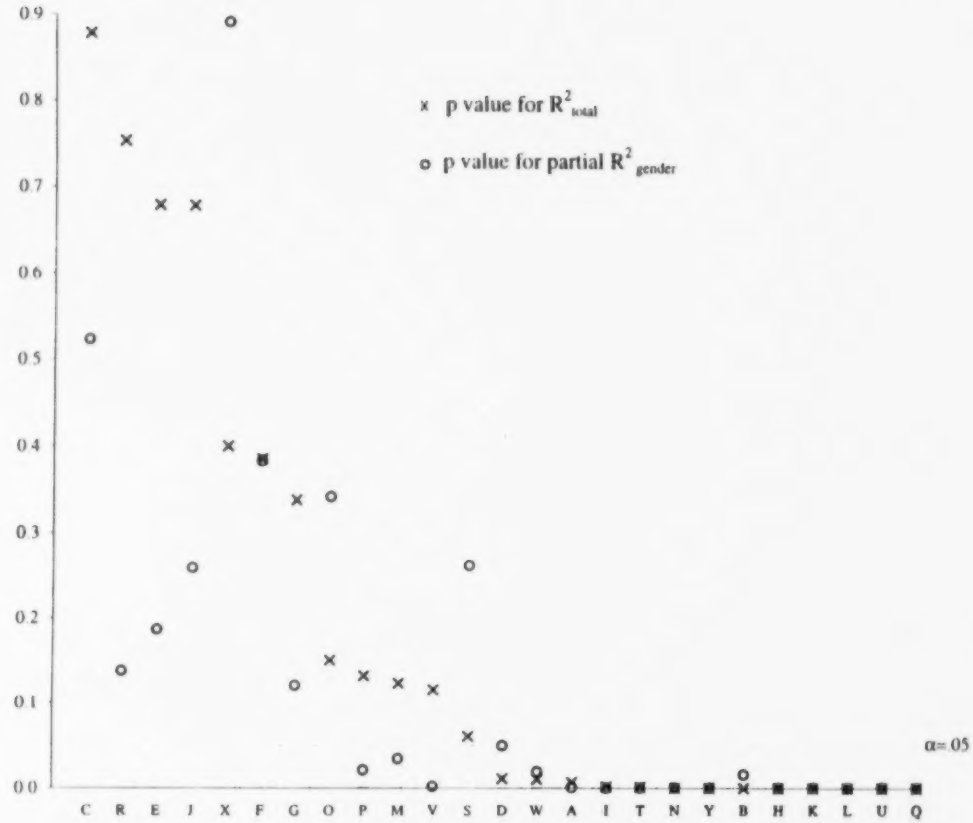

Figure 4. P values for R^2_{total} and partial R^2_{gender}

the student population (based on the average students' self-reported math grades, the countries/regions were categorized into either the more-able or the less-able group, relative to the rest of the countries/regions), and (b) the educational development level of the countries/regions (using parents' educational level as an index, two groups-- low or high were formed) (see Appendix A).

Estimation of common parameter

Homogeneous primary study results were integrated using variance-weighting to obtain the common parameter estimate (Shadish & Haddock, 1994). Standard error of the estimate and 95% confidence interval were also computed. In addition, general linear model procedure was applied to test the significance of common parameter.

Results and Discussions

The results of meta-analyses using various indicators, including the R^2_{total} from multiple regression model, the partial R^2 due to gender, the effect size for gender difference, the partial regression coefficient for gender, and p value, were discussed in this section. Overall outcomes of meta-analyses were summarized in Appendix C.

Combining Partial R^2s of Gender

Statistical considerations for combining partial R^2 due to gender and meta-analysis results are summarized in the following paragraphs.

Correction for bias. Despite the fact that the partial R^2s are biased due to small sample size (Hedges & Becker, 1990), to avoid negative values for the partial R^2s, the biases in this study were not corrected. Ranging from .002 to .016, the biases looked small anyway.

Estimation of variances. Since the R^2s of the 7-predictor model and the 6-predictor model were obtained from the same group of subjects, variance of the partial R^2 could be estimated by the following formula (Hedges & Becker, 1990):

$$V=([4\times R^2_{7\,predictors}\times(1-R^2_{7\,predictors})^2]/(n_m+n_f))$$

$$+\{[4\times R^2_{6\,predictors}\times(1-R^2_{6\,predictors})^2]/(n_m+n_f)\}$$

$$-2\times \mathrm{Cov}(R^2_{7\,predictors}, R^2_{6\,predictors})/(n_m+n_f) \quad ,$$

where n_m and n_f are the sample sizes for male and female groups respectively.

The appropriate estimate of the covariance term, $\mathrm{Cov}(R^2_{7\,predictors}, R^2_{6\,predictors})$, is considerably tedious and complicated. In previous research, the estimated covariance term was found to be trivial (B. J. Becker, personal communication, November, 1995). In this study, therefore, a judgmental decision was made to drop the term from the above formula to save time and labor.

Distribution of partial R^2. In large samples, the partial R^2 would have an approximately normal distribution. Approximate 95% confidence intervals were constructed for the observed R^2 of each primary study, based on the normal probability distribution. These are shown in Figure 5.

Homogeneity of the partial R^2s. The 95% confidence intervals of the partial R^2s from different studies appeared consistent. Most of the intervals were relatively narrow because of the small variance estimates for the partial R^2s.

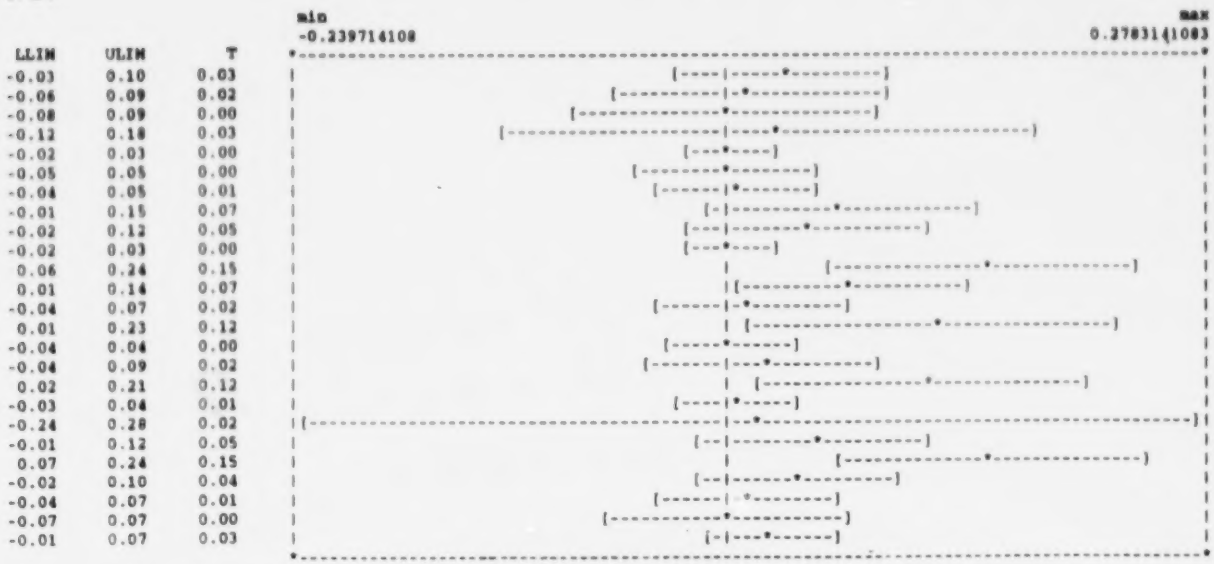

Figure 5. The 95% confidence intervals for the R^2_{chg}.

One-tailed confidence intervals might be more appropriate, because R^2s in the population can never be negative. Nevertheless, the approximate two-tailed intervals still provided a rough picture of the homogeneity of the primary results.

The hypotheses for the homogeneity test for the partial R^2s were:

$$H_0: \ \rho_1^2 = \rho_2^2 = \ .... \ = \rho_{25}^2 = \rho^2$$

and

$$H_a: \text{ At least one } \rho_i^2 \neq \rho^2, \ i = 1, 2, ..., 25,$$

where ρ^2 is the common population percent of variation attributable to gender alone. In words, the null hypothesis stated that all the partial ρ^2s from various study populations were equal, or that there was a common population ρ^2.

The results of the statistical test showed considerable agreement in the partial R^2s, with Q=36.007 (df=24, p=.055), which is not significant for α=.05. It was thus concluded that there was a common population partial ρ^2 across the countries/regions.

Although an outlier analysis identified two potential outliers, with standardized residuals greater than 2.5, the two extreme cases were not removed from the analysis. Since for international comparison studies such disparity was more likely to occur naturally, the primary results were kept intact. This decision was also justified by the results of the above homogeneity test.

Significance of common partial R^2. Using the variance-weighting method, the common partial R^2 was estimated to be .021, with a standard error of .005. This common R^2 estimate was found significant at α=.05 (Z=3.973>1.96). There appeared to be an overall non-zero partial R^2 for the gender variable, though the value is quite small.

The implication of this finding is that after the influences of the other

important variables were partialled out, across various countries/regions, gender still explains a statistically significant amount of variance in student's perceived gender differences in learning math. Specifically, female students thought female students would do better in math, whereas male student thought male students would do better.

Despite the statistical significance, the practical significance of the partial R^2 due to gender must be addressed. The magnitude of the impact of gender on students' perceptions about which gender group would do better on math might not be practically useful because the difference between the two gender groups was relatively small.

Combining R^2s from Multiple Regression Models

The meta-analysis results for the R^2_{total} from the multiple regression models (with seven independent variables) are summarized in the following paragraphs.

Estimation of variances. The variance of the model R^2 for each of the 25 studies was estimated as

$$V=[4\times R^2\times (1-R^2)^2]/(n_m+n_f) \ ,$$

where n_m was the sample size of the male student group, and n_f was the sample size of the female student group. The estimated variances were generally very small, ranging from .0001 to .0033 (i.e., standard errors ranging from .010 to .057).

Homogeneity of model R^2s. Though all were relatively small, the model R^2s from the 25 primary studies should quite a bit of variation. The distribution plot in Figure 2 displays the R^2 values. The values of these R^2s ranged from .009 to .220, with a mean of .076 and standard deviation of .059.

The null hypothesis for the homogeneity test for the model R^2s was that all the population model ρ^2s (from the 25 countries/regions) were equal to an overall common population ρ^2. The alternative hypothesis was that at least one

population ρ^2 was not equal to the overall common population ρ^2.

The results of the statistical test showed considerable disagreement in the model R^2s, with Q=75.948 (df=24, p≅0), which is significant for α=.05. It was thus concluded that the population ρ^2s were not homogeneous, and at least the ρ^2 of one country was different from the overall common ρ^2.

Outlier analysis. Potential outliers were identified empirically and four cases with standardized residuals larger than 2.5 were removed from the analysis. However, the Q statistic remained significant. To better summarize the global phenomena of gender difference in students' perceptions, a common estimate of population ρ^2 that summarized for as many countries/regions as possible is desired. Therefore, stricter criteria were applied to remove the most extreme cases until the remaining results were tested homogeneous.

After excluding a total of seven extreme cases (using 2.2 standardized residuals as the cut-off), test statistic Q became non-significant and the remaining 18 R^2s were found homogeneous. The seven outliers empirically identified in this model, however, may not be indeed outliers if some explanatory variables can be incorporated to the model to account for the differences between these outliers and the non-outliers.

Results of the homogeneity test for the remaining 18 R^2s indicated agreement with Q=23.958 (df=17, p=.121), not significant for α=.05. It was concluded that there was a common population model ρ^2 across the remaining 18 countries/regions.

Reasoning the outliers.

Other than the non-significant test result, it seems reasonable to combine the 18 country/region outcomes because (a) the number of outliers excluded was not too big, which lessened the risk of misidentifying outliers, (b) the criterion of 2.2 standard residuals was within the range of conventional criteria for identifying outliers, and (c) the values of the 18 remaining R^2s ranged from .015 to .220, with a mean of .066 and relatively

small standard deviation of .053. However, what actually separated the outliers from the non-outliers was not known because no clear patterns was found to link the seven outlier studies, or the 18 none-outlier studies.

It is shown that the result of homogeneity test could be reversed by arbitrarily defining outliers. Therefore, the cut-off criterion for identifying outliers should be chosen and justified with cautions. A careful review of extreme cases that incorporates information additional to standardized residuals is critical before making any decisions about outliers.

Estimating common R^2_{total}. The 18 homogeneous R^2s were combined by variance-weighting method to yield an estimate of common model R^2. The common population parameter was estimated to be .040, with a standard error of .005, which was relatively small. The 95% confidence interval around the weighted mean did not contain a value of zero, therefore, the common population R^2 estimate was concluded significant at α=.05. This confirms that there was an overall non-zero model R^2 for the 18 remaining countries.

The estimated common model R^2 (=.040) is about twice the size of the estimated common partial R^2 (=.021). It suggests a combined effect of the six independent variables, other than gender, in the full multiple regression model. It also shows that gender alone explains 50% of the variation in students' perceived gender differences in learning math. Nonetheless the common partial R^2 was estimated from all the 25 studies but the common model R^2 was estimated form 18 studies only.

Interpreting common model R^2. The significant common model R^2 estimate implies that the seven independent variables in the multiple regression model jointly accounted for a statistically significant amount of variance in students' perceived gender differences in learning mathematics for 18 countries/regions. Despite the statistical significance, the common model R^2 may not be useful in practice because of its small value.

Moderator Effect.

To avoid excluding outliers, heterogeneity of the model R^2s was studied by incorporating two potential moderators: (a) the general level of math

achievement of student population (level 1="less able" and level 2="more able"), and (b) the educational development level of the countries/regions (level 1="low" and level 2="high/medium").

The countries/regions were grouped and the between-groups heterogeneity and the within-group heterogeneity for the model R^2s were summarized in the table below:

Table 4

Summary of Heterogeneity for Model R^2s

Source of variation	Test statistic	Critical value	Decision	
Level of Math Achievement				
Overall	Q =75.948	χ^2_{24}	Reject H_0	
Between	Q_b= 2.526	χ^2_1	Retain H_0	
Within	Q_w=73.422	χ^2_{23}	Reject H_0	
"less-able"	Q_{w1}=34.996	χ^2_{11}	Reject H_0	(w-mean=.045, s.d.=.007)
"more-able"	Q_{w2}=38.426	χ^2_{12}	Reject H_0	(w-mean=.031, s.d.=.005)
Educational Development Level				
Overall	Q =65.308	χ^2_{22}	Reject H_0	
Between	Q_b= 6.971	χ^2_1	Reject H_0	
Within	Q_w=58.337	χ^2_{21}	Reject H_0	
"low"	Q_{w1}=33.187	χ^2_{12}	Reject H_0	(w-mean=.059, s.e.=.007)
"high/mdn."	Q_{w2}=25.150	χ^2_9	Reject H_0	(w-mean=.033, s.d.=.006)

Note. 1. For the between-groups tests, H_0: $\rho^2_{1\bullet} = \rho^2_{2\bullet}$ (i.e., no between-groups difference).

2. For the omnibus within-group test, H_0: $\rho^2_{11} = \rho^2_{12} = \ldots = \rho^2_{1m1} = \rho^2_{1\bullet}$ and $\rho^2_{21} = \rho^2_{22} = \ldots = \rho^2_{2m2} = \rho^2_{2\bullet}$ (i.e., no between-countries differences within each group), where m1 and m2 are the numbers of countries/regions in the two groups respectively.

3. Critical α=.05.

4. When Educational-development-level was coded, two studies were

excluded because of missing information. As a results, the Q statistic for this moderator is different from the Q for Math-achievement-level.

The between-studies heterogeneity accounted for by the two moderators were tested respectively for their significance at α=.05. No between-groups difference was found for math-achievement-level, suggesting that this moderator was not useful in grouping countries/regions into meaningful distinct groups. That is, math-achievement-level was not useful in explaining between-studies differences. The results of the within-group homogeneity tests further indicated the inadequacy of math-achievement-level, because the within-group population R^2s were not homogeneous in either groups.

The R^2s for the two groups formed by educational-development-level, nevertheless, appeared different. It suggested that the between-country(region) differences in R^2s was partly due to the differences in educational development level across countries/regions. The variance-weighted group means seemed to further indicate that the multiple regression model accounted for more variations in students' perceived gender differences on math learning for the group of educationally less developed countries than the model did for the group of more developed countries. Nonetheless, the within-group heterogeneity tests showed that the population R^2s were not homogeneous in either one of the two groups. The two estimated group means of population R^2s, therefore, should not be used to represent the average variance explained by the multiple regression model. In conclusion, educational-development-level is a better moderator than math-achievement-level, but its explanatory power is not sufficient for reasoning much of the between-studies differences. Despite the 11% of the between-countries variability explained by educational-development-level, a larger portion of between-countries heterogeneity (within the two groups) was unexplained.

Effect Sizes for Gender Differences

The unbiased effect sizes (see Figure 1) were tested for homogeneity. The unweighted mean of the effect sizes was about .350 and the unweighted standard deviation was about .268. The values ranged from -.141 to .845 and the distribution looked roughly normal.

Figure 6 showed that the 95% confidence intervals for the effect sizes might not be consistent, for some of the intervals looked quite far from the weighted mean effect size ($\bar{d}$=.326). In general, the intervals were relative narrow due to the small population variances estimated for individual studies.

Figure 6. The 95% confidence intervals for effect sizes.

The hypotheses tested for homogeneity were H_0: $\delta_1 = \delta_2 = ... = \delta_k = \delta$ and H_a: At least one $\delta_i \neq \delta$, where k=25, δ is the common population effect size, and i= 1, 2,...,25. In words, the null hypothesis stated that the population effect sizes for all of the 25 countries/regions were equal, and there was an common population effect size. It was found that the test statistic Q had a

BEST COPY AVAILABLE

value of 112.976 (greater than the critical value of χ^2_{24}), which was significant at α=.05. The null hypothesis was therefore rejected and it was concluded that the population effect sizes were not homogeneous across countries/regions.

The above test result seems to be different from the test result of partial R^2, which showed homogeneity among countries/regions. It is because the partial R^2 represents the effect due to gender uniquely, whereas the influence of the other important variables were not partialled out from the effect size. Without controlling for the effects of the other variables, the effect of gender is contaminated so the importance of gender (reflected by the magnitude of the effect size) decreased.

After removing seven outliers, with a cutoff criterion of standardized residual of 2.5, the remaining 18 population effect sizes appeared homogeneous at α=.05. The test statistic Q was 27.330, greater than the critical χ^2_{17}, and the p value was .053. The 95% confidence intervals in Figure 7 showed a pattern of consistency.

Figure 7. The remaining 95% confidence intervals for effect sizes, after

potential outliers were removed.

Significance of common population effect size. The variance-weighted common population effect size (δ) was estimated to be .320, with a relatively small standard error of .027. The 95% confidence interval around the weighted mean did not contain zero, and the significance test showed that $\hat{\delta}$ was significant at α=.05 ($p \cong 0$). Therefore, the null hypothesis that δ=0 was rejected. There was an overall non-zero population effect size across the 18 remaining countries/regions. Gender could explain a statistically significant amount of variance in student's perceived gender differences in math learning.

However, the gender effect found here was different from the gender effect found in the analysis of the partial R^2 due to gender. The partial-R^2 analysis modeled the effects due to some other important variables, while the effect-size analysis did not control for any possible effects of these variables. Similarly, meta analysis using partial $\hat{\beta}_{gender}$s would have different implications from the analysis using effect sizes. By partialing out the effects of other independent variables, given these independent variables are in fact important, the partial $\hat{\beta}_{gender}$s would be more powerful than the effect sizes.

Combining Partial $\hat{\beta}_{gender}$

Partial regression coefficients for gender ($\hat{\beta}_{gender}$s) were also meta-analyzed and the findings were compared to the meta-analysis results of the partial R^2.

Estimation of population variances. The population variances of $\hat{\beta}_{gender}$s for all the countries/regions were estimated by taking the squares of the standard errors of the $\hat{\beta}_{gender}$s, obtained from the SAS output for multiple regression analyses. The estimated variances were found relatively small (less than 1), compared to the magnitudes of the $\hat{\beta}_{gender}$s (see Table 2), except for one

study.

Homogeneity of $\hat{\beta}_{gender}$s. The $\hat{\beta}_{gender}$s had an unweighted mean of -1.565 and a relatively big standard deviation of 1.410. The weighted mean was -1.222, which had a standard error of .096). The 95% confidence interval plot in Figure 8 displayed an amount of variability among the $\hat{\beta}_{gender}$s for the 25 countries/regions.

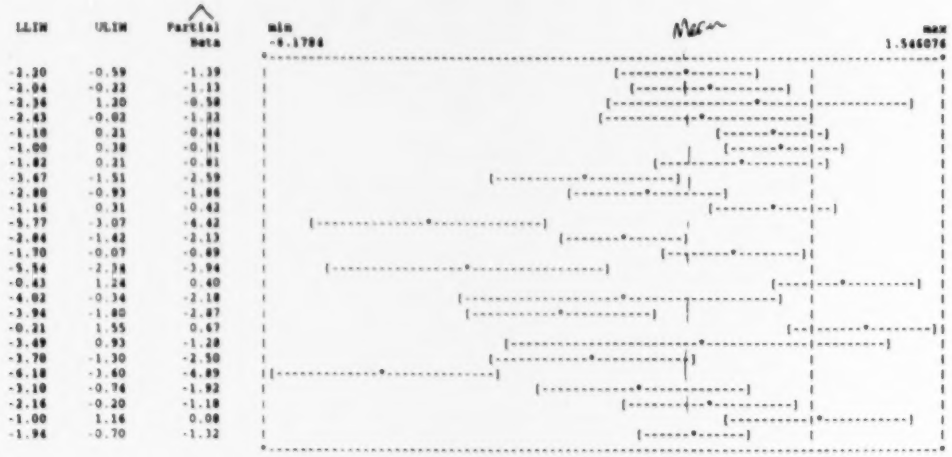

Figure 8. The 95% confidence intervals for partial $\hat{\beta}_{gender}$s.

The hypotheses of the homogeneity test for partial $\hat{\beta}_{gender}$s were analogous to those of the test for partial R^2s. The null hypothesis stated that all the population partial regression coefficients (β_{gender}s) were equal and there was a common population regression coefficient for gender. The alternative hypothesis stated that at least one β_{gender} was different from the common population regression coefficient. It was found that Q=150.377 (greater than the critical value of χ^2_{24}; p≅0) at α=.05. The $\hat{\beta}_{gender}$s appeared heterogeneous

across the 25 countries/regions (weighted mean=-1.222; standard error=.096), and at least one β_{gender} was different from the common population coefficient.

Outlier analysis. With a typical cut-off criterion (standardized residual greater than 2.5), eight extreme cases were excluded from the analysis, but the subsequent homogeneity test on the remaining 17 $\hat{\beta}_{gender}$ s still indicated between-studies heterogeneity (Q=31.753, p=.011). To avoid discarding primary studies, moderator effects were analyzed to explain the between-countries/regions heterogeneity in the $\hat{\beta}_{gender}$.

Moderator effects. The results of within-group heterogeneity were summarized in the table below:

Table 5

Summary of Heterogeneity for partial $\hat{\beta}_{gender}$ s

Source of variation	Test statistic	Critical value	Decision
Level of Math Achievement			
Overall	Q =150.377	χ^2_{24}	Reject H_0
Between	Q_B= 0.954	χ^2_1	Retain H_0
Within	Q_W=149.423	χ^2_{23}	Reject H_0
"less-able"	Q_{W1}=90.306	χ^2_{11}	Reject H_0 (w-mean=-1.358, s.d.=.169)
"more-able"	Q_{W2}=59.118	χ^2_{12}	Reject H_0 (w-mean=-1.157, s.d.=.117)
Educational Development Level			
Overall	Q =128.993	χ^2_{22}	Reject H_0
Between	Q_B= 5.708	χ^2_1	Reject H_0
Within	Q_W=123.285	χ^2_{21}	Reject H_0
"low"	Q_{W1}=90.588	χ^2_{12}	Reject H_0 (w-mean=-1.622, s.d.=.145)
"high/mdn."	Q_{W2}=32.697	χ^2_9	Reject H_0 (w-mean=-1.132, s.d.=.145)

Note. 1. For the between-groups tests, H_0: $\beta_{1\bullet} = \beta_{2\bullet}$.

2. For the omnibus within-group test, $H_0: \beta_{11} = \beta_{12} = \ldots = \beta_{1m1} = \beta_{1\bullet}$ and $\beta_{21} = \beta_{22} = \ldots = \beta_{2m2} = \beta_{2\bullet}$, where m1 and m2 are the numbers of countries/regions in the two groups respectively.

3. Critical α=.05.
4. When Educational-development-level was coded, two studies were excluded because of missing information. As a results, the Q statistic for this moderator is different from the Q for Math-achievement-level.

The moderator effects shown above are similar to the moderator effects found in the analysis for the model R^2 (see Table 4). Math-achievement-level was not useful in explaining between-countries differences in $\hat{\beta}_{gender}$s, whereas educational-development-level accounted for a small amount (about 4.4%) of the between-countries variation, suggesting that the gender difference in students' perceptions was somewhat dependent on the educational development level of the countries/regions. Although the variance-weighted group means seemingly implied that the group of educationally less developed countries/regions had larger gender difference in students' perceptions than the group of more developed countries/regions, the group means were misleading because the $\hat{\beta}_{gender}$s appeared different among countries within each of the groups. A large portion of the between-countries heterogeneity in $\hat{\beta}_{gender}$s was unexplained by this moderator. It was thus concluded that educational-development-level was not much better a moderator than the math-achievement-level.

Comparing results of partial $\hat{\beta}_{gender}$ and partial R^2_{gender}. Overall, the population partial regression coefficients of gender were not homogeneous across countries/regions. After controlling for the other variables, heterogeneity was found in students' perceptions of gender differences in learning mathematics across countries/regions. This conclusion is inconsistent with the conclusion reached by the analysis on partial R^2_{gender}, which showed homogeneity among country/region outcomes. However, although the partial R^2s appeared homogeneous, the p value (=.055) of the non-significant Q seemed marginal. The disparity between the results of the partial R^2s and the partial

$\hat{\beta}_{gender}$s may not be as serious as it is shown here.

One possible explanation for the inconsistent findings is that $\hat{\beta}_{gender}$s incorporate information on the direction of gender differences, whereas partial R^2s do not. Nevertheless, the reasoning may not be plausible because only three of the 25 studies have negative outcomes and they were all very small (see the group differences in Table 1, or the effect sizes in Figure 1).

The contradictory meta-analysis results have the following implications: (a) various statistics might have differential merits in meta-analysis, partly due to the differential approximations of their distributions such as the standard error estimates; (b) the covariance term dropped out of the estimation for the variance of the partial R^2 may in fact be important; and (c) the sensitivity of homogeneity test to the scales of various statistics needs to be addressed. The third implication is a bolder speculation based on the fact that the scale of partial R^2 was narrower than the scale of $\hat{\beta}_{gender}$ in this study. Naturally, the $\hat{\beta}_{gender}$ was more likely to have more variability, which might have contributed to the significant test statistic Q for $\hat{\beta}_{gender}$.

Combining p Values

Although the hypothesis tests for combining significance levels (ps) yield limited information about how the result of primary analysis vary from study to study (Becker, 1994), they are used when data other than significance levels are not available. Postulating a situation where significance levels are the only common information available for all countries/regions, this study analyzed various p values to compare the relative merits of different meta-analysis methods.

Hypothesis testing and interpretations. The p values for the multiple regression effects (model R^2) and the partial R^2s were studied. The distributions of these p values are displayed in Figure 4. Five commonly used methods were applied to summarize the p values (see Table 6). The hypotheses

Table 6

Methods and Results of Combining p Values

Test Method		Significance Level	Distribution of Test Statistic Critical Value	Test Statistics		Decision		Conclusion	
				For R^2 from multiple regression model	For R^2_{gender}	For R^2 from multiple regression model	For R^2_{gender}	For R^2 from multiple regression model	For R^2_{gender}
Sum of Z's Test		One-tailed α=.05	Z(0,1) Z^*=1.645	$\{\sum_{i=1}^{25} Z(P_i)\}/\sqrt{25}$ =9.273>Z^*	10.642>Z^*	Reject the H_0	Reject the H_0	At least one population R^2(I) is not zero.	At least one population R^2(I) is not zero.
Sum of Logs Test		α=.05	χ^2_{2k} (k=#of studies=25) *Critical* χ^2_{50} ~ 67.5	$-2\sum_{i=1}^{25} \log(p_i)$ = 226.917 > χ^2_{50}	252.456 > *Critical* χ^2_{50}	Reject the H_0	Reject the H_0	At least one population R^2(I) is not zero.	At least one population R^2(I) is not zero.
Logit Test		α=.05	$t_{(5k+4)}$ (k= # of studies=25) $t^*=t_{(129)}$~1.7	$-\sum_{i=1}^{25} \log(p_i/1-p_i)\{k\pi^2(5k+2)/3(5k+4)\}^{1/2}$ =11.738>t^*	=13.463>t^*	Reject the H_0	Reject the H_0	At least one population R^2(I) is not zero.	At least one population R^2(I) is not zero.
Minimum-p Test		α^*=.05 ; $\alpha=1-(1-\alpha^*)^{1/k}$ $=1-(1-.05)^{1/25}$ =.0021	[redacted]	Min(p1,p2....p25) =.0001 < α	Min(p1,p2....p25) =.0001 < α	Reject the H_0	Reject the H_0	At least one population R^2(I) is not zero.	At least one population R^2(I) is not zero.
Vote Count Method	(Conventional Vote Counting)	α=.05	Counts of significant & non-significant results Cut-off criterion=0.5	13/25=.52>.5	16/25=.64>.5	Reject the H_0	Reject the H_0	At least one population R^2(I) is not zero.	At least one population R^2(I) is not zero.

BEST COPY AVAILABLE

tested were $H_0: \rho_1^2 = \rho_2^2 = \ldots = \rho_{25}^2 = 0$, and H_a: At least one $\rho_i^2 \neq 0$ (or $\rho_i^2 > 0$), where i=1,2,...,25.

For the p values of the model R^2, the null hypothesis stated that all the 25 population model ρ^2 were equal to zero. That is, none of the 25 regression models accounted for a statistically significant amount of variance in students' perceived gender difference. The alternative hypothesis was that at least one population model ρ^2 was significant, that is, at least one regression model was useful in explaining the variation in the dependent variable. Exact number of populations with useful models remains unknown from this analysis. The explanatory power of the model(s) is also not clear. In addition, it was hard to know in what study populations the regression model fits.

For the p values of the partial R^2, the null hypothesis stated that none of the 25 population partial ρ^2s was useful in explaining the variation in students' perceived gender difference. The alternative hypothesis was that for at least one country/region gender was useful in explaining the variation in the dependent variable. Exact number of populations with important gender effect remains unknown from this analysis. The magnitude of the gender effect is not clear, either. It is also hard to know in what study populations the gender effect is credible.

Comparisons among methods. The five test methods used were the sum of Zs test, the sum of logs test, the Logit test, the minimum-p test, and conventional vote counting. Table 6 reports the observed test statistics, the critical values (or significance levels), the distributions of the test statistics, as well as the decisions and conclusions reached by these methods. It was found that the test results were consistent over various methods, all suggesting the rejection of the null hypotheses. Every summary indicates that at least in one country/region the population ρ^2 was not zero.

Counting Positive (or Significant Positive) Results

The non-parametric sign-test method (Bushman, 1994) was used to summarize

the effect sizes from various countries/regions (the R^2s were not used because they were non-directional). As an alternative vote count method, the sign test counts either the number of studies with positive results or the number of studies with significant positive results. The results of the sign tests are summarized in Table 7.

This study first used the sign test to examine whether the effect sizes from the 25 independent studies (countries/regions) were all zero. Let π_p be the proportion of positive results in the population, given the fact that the probability of getting a positive result is 0.5 when the effect sizes were all zero, the hypotheses tested were H_0: π_p=0.5 and H_a: π_p>0.5. The null hypothesis was based on the cumulative binomial distribution (Berry & Lindgren, 1990). Among the 25 effect sizes found in this study (see Figure 1), 22 were positive and only 3 were negative. Therefore, an estimator of π was p=0.88 (=22/25). The sign-test result of the above null hypothesis suggested that the proportion of positive results in the population was greater than 0.5, which corresponded to the alternative hypothesis that not all the 25 effect sizes were zero. Therefore, at least one effect size appeared to be non-zero.

Then, the sign test was used to test a second pair of hypotheses--H_0:π_s=0.05 and H_a:π_s>0.05. The null hypothesis stated that the proportion of <u>significant positive</u> effect sizes in the population was no more than the expected .05 (or there were no more significant effect sizes across studies than expected). Because the observed proportion of statistically significant positive effects was large, the sign-test probability was smaller than the critical value of α=.05 (as shown in Table 7). The null hypothesis was rejected and it was concluded that not all the effect sizes appeared to be zero. At least in one country/region the effect size was statistically significant and positive.

The disadvantages of the sign test are (a) it does not take into account sample size, and (b) it does not offer an estimate for the overall effect size.

Table 7

Results of Sign Tests for Effect Sizes

Test Method		Significance Level	Distribution of Test Statistic	Hypotheses	Test Statistic	Decision	Conclusion
Vote Count Method	Sign Test for Positive Results	$\alpha=.05$	Cumulative binomial probability distribution	H_0: $\pi=.5$ H_a: $\pi>.5$	$\sum_{i=22}^{25}\binom{25}{i}p^i(1-p)^{25-i}$ $=0.0001<0.05$	Reject the H_0	Not all the 25 population effect sizes are all zero.
	Sign Test for Significant Positive Results	$\alpha=.05$	Cumulative binomial probability distribution	H_0: $\pi=.05$ H_a: $\pi>.05$	$\sum_{i=15}^{25}\binom{25}{i}p^i(1-p)^{25-i}$ $=.0000<0.05$	Reject the H_0	Not all the 25 population effect sizes are all zero.

These are common drawbacks of conventional vote-count methods. As a consequence, the vote counting procedures used in this study have lower power than the other methods (Bushman, 1994).

Issues and Suggestions

Several meta-analysis issues emerged from this study, including the non-directional nature of R^2, the explanatory effect of moderators, and the selection of meta-analysis indicator from a variety of available summary statistics, are elaborated in this section. Suggestions are provided for future cross-national meta-analysis as well. The overall advantages of meta-analysis are summarized in the light of improving the validity of international comparison studies.

Combining R^2

As Hedges and Olkin (1985) commented, R^2 is probably not best suited for combination across studies. One problem is its inherent non-directional nature. Similar values of R^2 may be obtained with substantively different results in various studies. For bivariate relationships, for instance, correlation coefficients with the same magnitudes but different signs result in R^2s of identical values.

For the regression model R^2s analyzed in this study, the lack of direction was not relevant because the R^2 from a multiple regression was used to determine the overall explanatory power of the entire model. For the partial R^2s due to gender, however, the use of the non-directional R^2 was problematic. The corresponding partial regression coefficients for gender revealed that the gender effects in fact had different directions for different countries/regions.

The few studies that had results different in direction from the rest should be carefully reviewed to determine whether their results truly reflected inconsistent phenomena across countries/regions, or they were influenced by hidden moderator(s), or they were simply anomalies due careless study designs. Other than theoretical considerations, additional information from other

empirical sources may be needed for the review. If it is determined that these studies were influenced by careless study designs, they should be removed from the analysis so the common parameter estimates would not be contaminated. If it is found that these studies spoke for real phenomena of students' perceived gender differences on math learning, despite the fact that their results looked different from the rest, it makes sense to include these studies when common population parameters are estimated. However, moderator effects should be exploited to explain why the study results had different directions.

Another problem that may prevent the use of R^2 in meta-analysis is that R^2 may be sensitive to the definition of groups or patterns of the values of independent variables (Hedges and Olkin, 1985). In other words, the size of R^2 depends on not only the relationship between the predictor(s) and the outcome variable, but also the particular value-ranges chosen for the independent variables. Across various primary studies, even when the same independent variables are used, the definitions of the scales of independent variables may vary from study to study. In this study, such problem of the R^2 did not exist because the scale for each of the independent variables in TIMSS survey questionnaire was invariant across all countries/regions.

Exploiting Explanatory Moderators

It is shown in this study that moderators have great potentials in explaining between-countries differences in primary study outcomes. This meta-analytic approach went beyond the flat summary of conventional review method and proposed sensible models to explain why the country outcomes varied. The models were further tested for their statistical significance. Although the two moderators used in this study were generally not satisfactory in accounting for much of the between-countries differences, if some other explanatory moderators could be found and incorporated to the analysis, the reasons for between-countries differences would be identified and verified.

One possible cause of the disappointing moderator effects in this study is that the two moderators are not sufficiently gender-specific. In future studies, potential moderators that are sound in theory or more gender-specific

should be exploited to better explain the cross-national variation in the gender difference in students' perceptions about whether girls or boys will do better on math.

Relative Merits of Meta-analysis Indicators

This study showed that various meta-analysis indicators, such as partial R^2_{gender} and partial $\hat{\beta}_{gender}$, may yield inconsistent synthesis results, suggesting differential merits of these indicators. It is therefore important to select an indicator that relatively best suits the interest of a meta-analysis. Or, multiple indicators can be used and their results can be compared to cast insights on the applicability and plausibility of various statistics.

In this study, because the homogeneity found in the partial R^2s seemed marginal and the heterogeneity found in the partial $\hat{\beta}_{gender}$s was consistent with the findings based on the effect sizes, the meta-analysis result yielded by the $\hat{\beta}_{gender}$ seems plausible. In addition, because R^2 do not take into account the information on the directions of the gender differences, $\hat{\beta}_{gender}$ is a relatively better indicator for synthesizing cross-national study outcomes on the gender differences in students' perceptions. Furthermore, the partial $\hat{\beta}_{gender}$ is better than the effect size d because it represents unique effect due to gender, by controlling for the other important variables. Overall, the theoretical advantages of partial $\hat{\beta}_{gender}$ fit the purpose of this international study best, and the findings based on this indicator seem plausible.

An extension of the issue on meta-analysis indicator is the compatibility of primary study outcomes based on different statistics. The issue is especially important when transformation is required to arrive at comparable statistics for various primary studies. For instance, if correlation coefficients (rs) are to be summarized to depict a common bivariate relationship but some rs are not directly available from some of the studies, then these missing rs can be retrieved via data transformation, such as (a)

taking the square root of R^2 from simple regression ($|r| = |\sqrt{R^2}|$) and attaching an appropriate sign of direction, or (b) transforming an effect size to an r. While obtaining a desirable meta-analysis indicator for all of the studies by data transformation, special attentions should be paid to make sure the information from different studies are compatible.

Advantages of Meta-Analysis for International Studies

Overall, the meta-analytic techniques applied in this study are satisfactory in analyzing the TIMSS data, indicating great potentials of meta-analysis in improving the validity of international comparison studies. It is shown that meta-analysis is not only useful in integrating homogeneous country results, but it is also capable of detecting substantial differences in country outcomes and effective in offering strategies to deal with such situation.

By outlier analysis and the study of moderator effects, meta-analysis is likely to provide explanations for inconsistent country findings. By grouping countries into meaningful homogeneous sub-groups and estimating common parameters for the countries within each of the sub-groups, meta-analysis will present more realistic and statistically sound pictures of global phenomena in education, such as students' perceptions of gender differences on math learning. Hence, validity of international comparisons can be improved upon a qualitative basis.

Due to the limitations in the nature of the TIMSS data and the study design, the conclusions and implications reached in this study should be carefully interpreted and generalized to appropriate populations and occasions.

Reference

Bandura, A. (1982). Self-efficacy mechanism in human agency. American Psychologist, 37, 122-147.

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy and intrinsic interest through proximal self-motivation. Journal of Personality and Social Psychology, 41, 568-598.

Becker, B. J. (1994). Combining significance levels. In Harris Cooper & Larry V. Hedges (Eds.), The handbook of research synthesis (pp.215-230). New York: Russell Sage Foundation.

Berry, D. A., & Lindgren, B. W. (1990). Statistics: Theory and methods Belmont, CA: Brooks/Cole.

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In Harris Cooper & Larry V. Hedges (Eds.), The handbook of research synthesis (pp.193-214). New York: Russell Sage Foundation.

Eagly, A. H., & Wood, W. (1994). Using research syntheses to plan future research. In Harris Cooper & Larry V. Hedges (Eds.), The handbook of research synthesis (pp.485-500). New York: Russell Sage Foundation.

Fennema, E. L. (1974). Mathematics learning and the sexes: A review. Journal for Research in Mathematics Education, 5, 126-139.

Fennema, E. L., & Sherman, J. A. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. American Educational Research Journal, 14, 51-71.

Hedges, L. V. (1994). Statistical considerations. In Harris Cooper & Larry V. Hedges (Eds.), The handbook of research synthesis (pp.29-38). New York: Russell Sage Foundation.

Hedges, L. V., & Becker, B. J. (1990). Detecting and measuring improvements in validity (Contract No. DAAL03-86-D-0001). Navy Personnel Research and Development Center.

Hedges, L. V., & Olkin I. (1985). Statistical methods for meta-analysis Boston: Academic Press.

Hermans, H. J. M. (1970). A questionnaire measure of achievement motivation. Journal of Applied Psychology, 54, 353-363.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavior research. New York: Holt, Rinehart & Winston.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Mayer, R. E. (1987). Educational psychology: A cognitive approach. Boston: Little, Brown and Company

Norwich, B. (1986). Assessing perceived self-efficacy in relation to mathematics tasks: A study of the reliability and validity of assessment. British Journal of Educational Psychology, 56, 180-189.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. Journal of Educational Psychology, 73, 93-105.

Schunk, D. H. (1988, April). Perceived self-efficacy and related social cognitive processes as independent variables of student academic performance. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In Harris Cooper & Larry V. Hedges (Eds.), The handbook of research synthesis (pp.261-284). New York: Russell Sage Foundation.

Shavelson, R. J. (1988). Statistical reasoning for the behavioral sciences. Needham Heights, MA: Allyn and Bacon.

Appendix A

Coding Schemes and Outcomes of Moderators

Moderator / Data / Country	Level of Math Achievement		Educational Development Level	
	Raw Score	Category	Raw Score	Category
A	3.139	H	0.478	M
B	2.841	L	0.059	L
C	2.798	L	0.468	M
D	3.197	H	0.567	M
E	3.016	H	0.382	M
F	3.061	H	0.161	L
G	3.056	H	0.285	L
H	3.279	H	0.144	L
I	3.089	H	0.350	M
J	3.103	H	unknown	—
K	3.126	H	0.192	L
L	3.140	H	0.393	M
M	3.031	H	0.115	L
N	2.971	L	0.139	L
O	2.229	L	unknown	—
P	2.596	L	0.438	M
Q	2.901	L	0.103	L
R	2.926	L	0.413	M
S	2.917	L	0.400	M
T	2.852	L	0.100	L
U	2.867	L	0.236	L
V	2.751	L	0.084	L
W	2.725	L	0.227	L
X	3.090	H	0.199	L
Y	3.230	H	0.557	M

Note 1. For moderator Math-achievement-level, H="more able" and L="less able"; range=(1,4); average = 3;
if (raw score<=3) then L; if (raw score>3) then H.

2. For moderator Educational-development-level, L="Low", M="Medium", and H="High"; range=(0,1);
if (raw score<1/3) then L, if (1/3=<raw score<2/3) then M, if (raw score>=2/3) then H.

Appendix B

Independent Variables and Component Questionnaire-items

° Student's gender

° Student's general educational aspiration

° Student's achievement attribution (external or intrinsic):

To do well in math you need--
- lots of natural talent
- good luck
- lots of hard work studying the subject
- to pay attention in class
- to memorize the textbook or notes

° Student's general preference of mathematics:

How much do you like mathematics--
- Do you enjoy studying mathematics?
- Is the study of mathematics important?
- Is mathematics a hard subject?
- Is mathematics boring?
- Is mathematics important to everyone's life?
- Would you like a job that involved using mathematics?

° Parents' education:

- Father's education
- Mother's education

° Student's overall perceived expectations on mathematics learning:

- My parents think it is important for me to do well in mathematics classes
- Most of my friends think it is important for me to do well in mathematics classes
- I think it is important for me to do well in mathematics classes

° Student's mathematics self-efficacy :

- How well do you usually do in mathematics?

Appendix C
Summary of Meta-analysis Results

Primary Study Outcome	Test statistic (Q)	df	Significance level (p)	Average parameter estimate (variance-weighted)	(Std. error)	Significance
partial R^2_{gender}	36.007	24	0.055 (n.s.)	0.021	(.005)	*
R^2_{total}	75.948	24	$\equiv 0$	0.036	(.004)	n.t.
R^2_{total} (*w/ o 7 outliers*)	23.958	17	0.121 (n.s.)	0.040	(.005)	*
$\hat{\beta}_{gender}$	150.377	24	$\equiv 0$	-1.222	(.096)	n.t.
$\hat{\beta}_{gender}$ (*w/ o 8 outliers*)	31.753	16	0.011	-1.139	(.118)	n.t.
d	112.976	24	$\equiv 0$	0.326	(.024)	n.t.
d (*w/ o 7 outliers*)	27.330	17	0.053 (n.s.)	0.320	(.027)	*

Note: *-- Significant at α=.05

n.t.-- Not tested. Because the primary study outcomes are heterogeneous, indicated by the significant Q, the average parameter estimate does not reflect gender difference on students' perceptions across countries/regions.